基于DenseNet的天体光谱分类方法

王奇勋[12] 赵刚[12] 范舟[1]

(1 中国科学院光学天文重点实验室国家天文台 北京 100101)

(2 中国科学院大学天文与空间科学学院 北京 100049)

摘 要：天体光谱数据的智能处理正由传统机器学习方法逐步转向深度学习，主要采用基于计算机视觉的技术手段。本文基于在计算机视觉领域广泛应用的 DenseNet 网络结构，针对光谱数据进行修改，建立了适用于光谱数据的一维卷积神经网络模型解决天体光谱数据分类任务。在验证数据集上，恒星、星系、类星体的 F1 分数达到了为 0.9987、0.9127、0.9147，高于传统神经网络。光谱分类关注区域的可视化结果表明，本文模型可以学习到各类天体对应的特征谱线，具有较强的可解释性。本文的方法被用于阿里云天池天文数据挖掘大赛——天体光谱智能分类，并在 843 支参赛队伍的 3 次数据评比中获得了 2 次第一、1 次第三的成绩，证明了该模型在保证分类精度的同时具有极强的鲁棒性、泛化性，适用于光谱的自动分类。

关键词：卷积神经网络；光谱分类；数据挖掘

1. 研究背景与意义

随着科学技术的发展和观测设备不断升级，天文数据呈现爆炸式的增长。人工智能（AI）技术能够辅助天文学家们处理分析海量天文数据，发现新的特殊天体和物理规律。天体光谱数据的智能处理正由传统机器学习方法逐步转向深度学习[1,2,3]，主要采用基于计算机视觉的技术手段。参考文献[1]提出了使用 5 层卷积神经网络估计大气参数的方法。参考文献[2]提出使用自编码算法的神经网络对斯隆数字巡天（SDSS）光谱进行恒星大气物理参数的估计。参考文献[3]提出使用深度神经网络模型并构造分类器对光谱进行分类。深度学习方法较机器学习在处理天体光谱数据上的精度、鲁棒性和泛化性都有明显提升。

基金项目：国家自然科学基金（11390371）资助.

作者简介：王奇勋，男，博士，研究方向：计算机视觉，光谱分类.Email:wangqixun@nao.cas.cn

郭守敬望远镜（LAMOST，大天区面积多目标光纤光谱天文望远镜）是一架新类型的大视场兼备大口径望远镜，在大规模光学光谱观测和大视场天文学研究方面，居于国际领先的地位。LAMOST 是世界上光谱获取率最高的望远镜，每个观测夜晚能采集万余条光谱。截止 LAMOST DATA RELEASE 5 v1，已产生 900 万以上条光谱。光谱类别的划分是所有天文研究的前提，正确的分类可以减少天文学家对数据的筛选、清洗工作，同时可以提高 LAMOST 数据的使用效率。光谱自动分类是从上千维的光谱数据中选择和提取对分类识别最有效的特征来构建特征空间。天体光谱与其类别之间的关系是高度非线性的，实际观测中又存在大量观测来自仪器、天气方面的噪声。传统的机器学习与模板匹配方法在非线性关系的探索上表现不佳，而深度学习在非线性关系探索和表示方面有着天然的优势，理论上高深度的神经网络可以拟合任意复杂的关系函数。

因此，本文基于 LAMOST 光谱数据特点对 DenseNet[4]网络结构进行优化，并利用优化后的模型结构对光谱数据进行分类训练与检测，提出了基于 DenseNet 的 LAMOST 光谱自动分类处理方法。本方法被用于阿里云天池天文数据挖掘大赛——天体光谱智能分类，并在 843 支参赛队伍的 3 次数据评比中获得了 2 次第一、1 次第三的成绩，证明了该模型在保证分类精度的同时具有极强的鲁棒性、泛化性，适用于光谱的自动分类。

2. 数据集以及类别

本文数据集选取自国家天文与阿里云天池主办的“天文数据挖掘大赛”发布的 LAMOST 光谱数据[1]，共 449384 条。随机抽取其中 89877 条光谱作为验证集数据，其余作为原始的训练数据。海量不同但是具有同一类别的光谱数据，可以显著的提高模型的泛化能力，同时使模型在验证数据集上拥有很强的适应性。

LAMOST 数据集中的每一条光谱提供了 3690—9100 埃的波长范围内的一系列辐射强度值。该竞赛是纯粹的数据挖掘比赛，光谱数据不含波长信息，但所有光谱数据的波段区间和采样点相同，采样点个数都是 2600 个。

在目前的 LAMOST 巡天数据发布中，光谱主要被分为恒星（STAR）、星系（GALAXY）、类星体（QSO）和未知（UNKNOWN）天体四大类，本文仅对恒星、星系和类星体三类天体光谱进行分类。

3. 分类模型

[1]天文数据挖掘大赛: https://tianchi.aliyun.com/competition/entrance/231646/information

天体光谱数据是天体的点源光经过色散形成的、分布在不同波长下流量强度的序列。根据其吸收线、发射线的位置、强弱、宽度等性质，天文学家可以判断该天体的所属类别。可以说天体光谱的在一维坐标系下的“样子”决定了它的类别，这种“样子”决定“类别”的任务天然适合于卷积神经网络。传统的卷积神经网络预训练模型都是针对二维图像数据构建并训练得到的，并不适用于一维天体光谱数据。为此，我们基于在计算机视觉领域广泛应用的 DenseNet 网络结构，针对光谱数据进行修改，建立了适用于光谱数据的一维卷积神经网络模型解决天体光谱数据分类任务。网络模型结构如图 1。网络模型通过 N+1 个 Dense Block 结构依次连接，在 Dense Block 中采用密集型连接结构解决深层网络带来的梯度弥散问题。

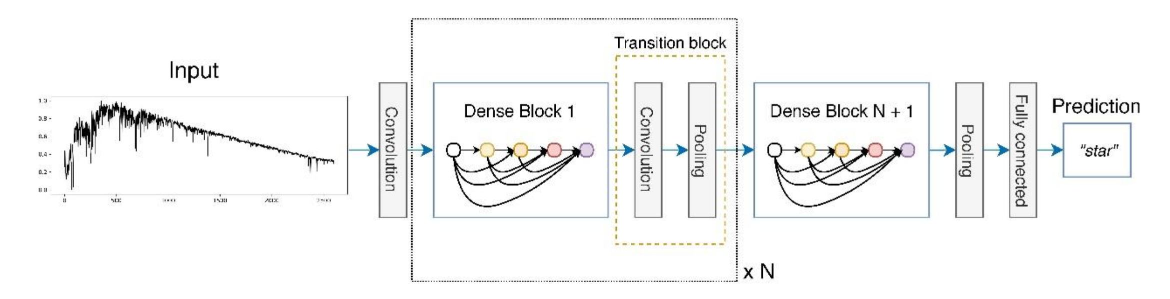


图 1 本文使用的深度神经网络结构。蓝色框中是 Dense Block 结构。黄色框中是 Transition Block 结构。输入数据依次通过 N+1 个 Dense Block 并得到新的特征，最终输出对输入光谱的预测。

Fig.1 The structure of deep neural network in this paper. The blue box represented the Dense Block structure. The yellow box represented the Transition Block structure. The input data flowed through N+1 Dense Block in turn to generate new feature maps, and subsequently the prediction result was generated.

数据流的处理过程为，归一化后的光谱数据经过卷积层后进入 Dense Block 结构，经过 N+1 次 Dense Block 的特征提取得到高维特征图，随后对每个高维特征图进行全局平均，平均后的值再经过全连接层输出该光谱所属类别的概率预测值。

除最后一个 Dense Block 外，所有 Dense Block 后都会连接一个卷积层和一个池化层。二者的目的都是为了减少计算量，卷积层从数据维度上减低计算量，池化层从数据尺度上降低计算量，这两层的连接记为 Transition Block。但是池化层通过将两个相邻的数据点平均成一个数据点，将原始数据的尺度变成了原来的一半，使得不同尺度下 Dense Block 间不能再进行密集型的连接。为了进一步增强不同尺度下特征的联系和重用，我们对原 DenseNet 结构进行改进，在两个 Transition Block 间增加了额外的 Transition Block 跨层连接，如图 2 所示。

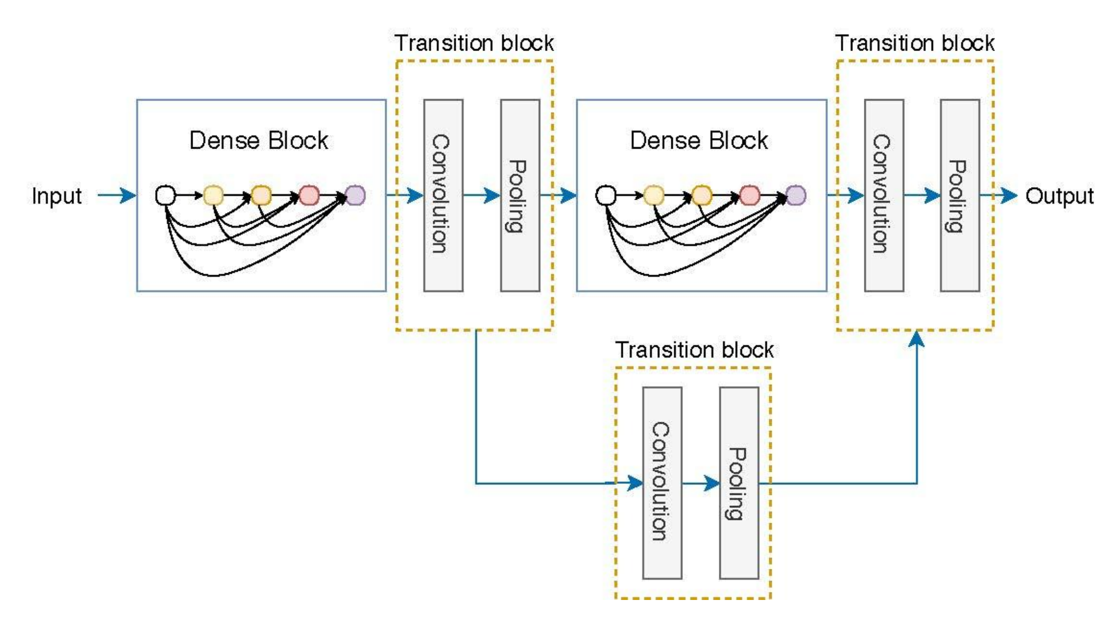


图 2 两个 Transition Block 间额外的 Transition Block。额外的 Transition Block 进一步增强了不同尺度下特征的联系和重用。

Fig.2 Additional transition blocks between two transition blocks. Additional transition block enhances the connection and reuse of feature maps at different scales.

Dense Block 通过 Conv Block 密集型连接构成，如图 3。密集型连接使得在同一个 Dense Block 中，所有卷积层的输入来源于前面所有层的输出，加强了不同特征间的联系，也缓解了深度神经网络在训练时会遇到的梯度弥散问题。在本试验中，我们使用 2 个 Conv Block 连接。更少的 Conv Block 使分类精度下降，更多的 Conv Block 不仅不会带来更高的精度提升，而且会使计算量急剧增加。

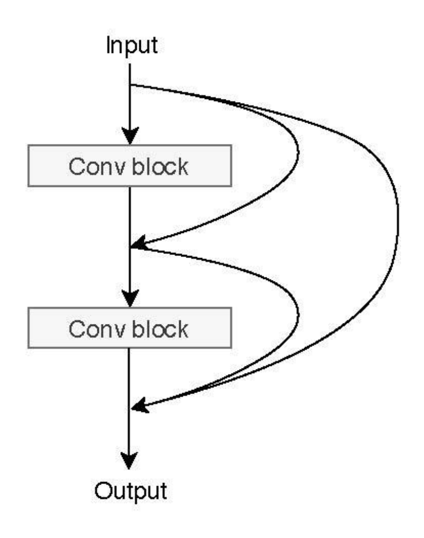


图 3 Dense Block

Fig.3 Dense Block

Conv Block 作为具体的特征提取结构，需要针对输入数据进行针对性的构建。本文使

用的天体光谱数据具有发射线、吸收线等物理特征，由于这些元素的谱线宽度、位置不确定，我们使用多个卷积步长（1、11、25、41）的卷积层分别对输入的数据进行卷积，将得到的不同卷积尺度下的卷积结果合并到一起后，使用卷积步长为 1 的卷积层进行不同卷积尺度下的整合，如图 4，使得神经网络模型可以对元素谱线进行更好的适应与学习。其中批归一化（Batch normalization）[5] —特征重标定结构（Squeeze-and-Excitation block）[6] —修正线性单元（Rectified Linear Unit）[7]依次连接的处理加在了每一次卷积层前。批归一化和修正线性单元都是为了进一步缓解深度神经网络在训练时遇到的梯度弥散文题。

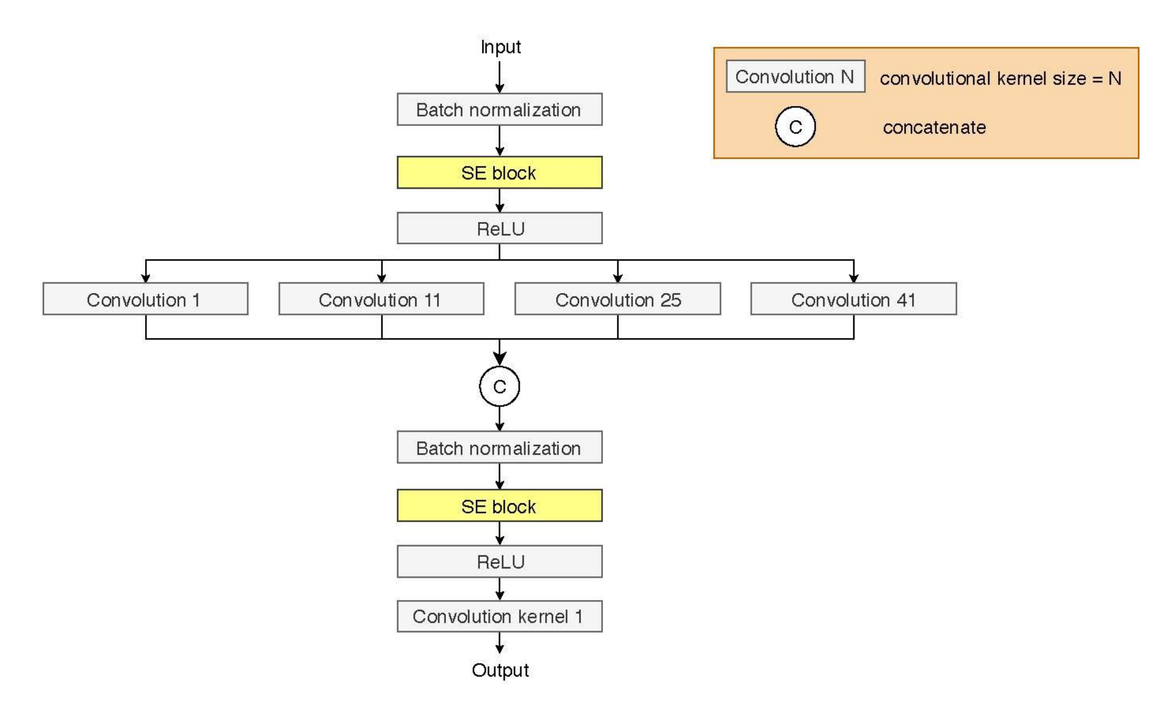


图 4 conv block。合并多个卷积尺度（1、11、25、41）下的特征图，并抽取新的特征。

Fig.4 Conv block. The new feature maps are generated by extracting merged feature maps which convoluted from multiple scales (1, 11, 25 and 41).

卷积层前的特征重标定结构会使神经网络模型对产生的新特征进行再次选择。这个“选择”的过程通过对原始特征赋权重实现。通过学习的方式来自动获取到每个特征的重要程度，然后依照这个重要程度去提升有用的特征，并抑制对光谱分类任务用处不大的特征。具体的实现方式如图 5。原始特征图通过全局平局池化转换为点向量，随后通过全连接层，输出每个特征图对应的权重，最后将原始特征图与该权重相乘，输出加权后的新特征图。

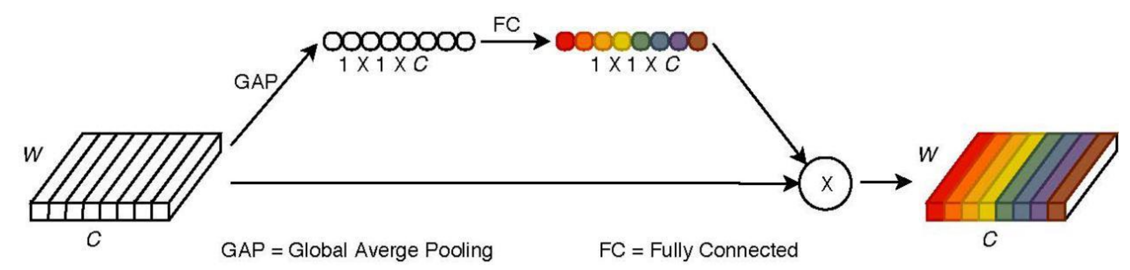


图 5 SE block。原始特征图通过全局平均池化和全连接后产生对应的权重。原始特征与该权重相乘后输出新的特征图。

Fig.5 SE block. The feature maps generated corresponding weight through global average pooling and full connection. The new feature maps were generated by multiplying the original feature maps and the corresponding weight.

4. 分类网络模型训练

4.1 数据归一化

由于不同天体视星等、曝光时间不同，观测到的光谱流量差异巨大。为使分类网络模型更加鲁棒，能够适应不同流量差异，更快的学习各天体类别间的差异，我们对分类网络的输入数据进行归一化处理。如公式（1）操作：

$$F = \frac{F - F_{min}}{F_{max} - F_{min}} \#(1)$$

其中 F 为原始光谱流量数据，F_{min}和F_{max}为原始光谱流量数据的最小值和最大值。光谱归一化前后如图 6 所示，左边的是 3 条光谱归一化前的流量可视化展示，右边的图为可视化后的流量展示。归一化前不同光谱的流量不在同一数量级，归一化后不同光谱的流量统一。归一化有利于卷积神经网络更加快速的学习不同天体光谱间的特征差异，使神经网络关注的重点不包含流量因素。归一化使神经网络的训练速度加快、精度提升。

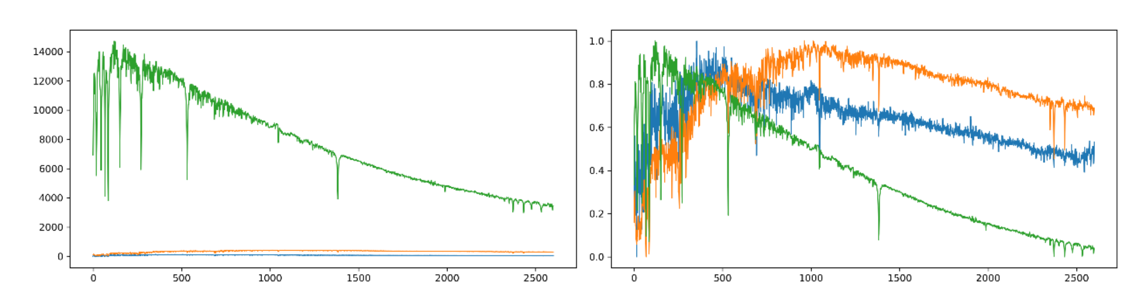


图 6 归一化前后对比。左侧为归一化前的光谱，右侧为归一化后的光谱

Fig.6 The specturm data before and after normalization. The left is the pre-normalized spectrum and the right is the normalized spectrum.

4.2 数据增强

实验数据中恒星、星系和类星体三类天体所占比例为 98.5:1.2:0.3，数据极度不平衡，恒星类型远多于其他两类，所占比例如图 7 所示。不平衡的数据会使卷积神经网络的分类效果下降，数据扩充是非常必要的。我们通过在星系、类星体数据上增加具有具体物理意义的噪声进行过采样扩充数据，从而增加星系、类星体所占比例；在恒星类型数据上进行欠采样减少该类别所占比例。通过数据扩充，恒星、星系和类星体三类天体所占比例平衡到 1:1:1。

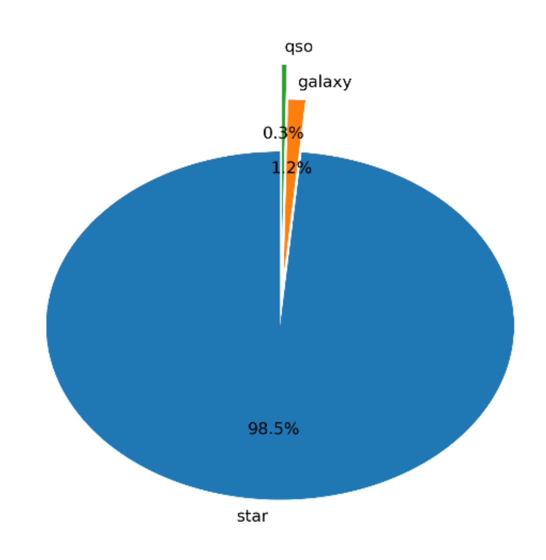


图 7 原始数据中恒星、星系和类星体数量占比

Fig.7 The proportion of stars, galaxies and quasars in raw data.

在星系、类星体数据增加的噪声是模拟天体光谱在产生过程中不可避免的噪声情况。本实验中，我们模拟了三种噪声：CCD 电子噪声、CCD 坏像元和宇宙高能射线。

4.2.1 CCD 电子噪声

天体光谱最终的接收端是 CCD 相机。来自遥远天体的光经过色散后被 CCD 相机接收，通过光电效应产生电子，光子信息以电子的形式被记录下来。由于 CCD 相机在工作时产生的电子也会被自身接收，从而天体光谱产生的电子与 CCD 自身的电子会同时被 CCD 当做光子信息记录，由此产生 CCD 电子噪声。CCD 工作带来的电子噪声在信噪比越低时越明显。我们通过生成一个 2600 列的高斯随机数模拟 CCD 电子噪声，随后加在归一化后的天体光谱上进行过采样扩充数据。增加模拟 CCD 电子噪声前、后的光谱可视化如图 8。第 1、2、3 行展示了随机选取的恒星、星系、类星体天体类型光谱，第 1、2 列分别展示了对应的天体类型的原始光谱和模拟了 CCD 电子噪声后的光谱。

4.2.2 CCD 坏像元

CCD 相机的接受端难免出现故障，导致任何强度的光子信息都接受不到。在光谱上的表现为某一波段的流量强度为 0。虽然实际情况下，这种存在坏像元影响的光谱极少，但为了扩充数据，并使分类模型具有更强的泛化性和鲁棒性，我们通过将归一化后的光谱中某一随机波段随机长度的流量置为 0 去模仿 CCD 坏像元。增加模拟 CCD 坏像元前后的光谱可视化如图 8。第 1、2、3 行展示了随机选取的恒星、星系、类星体天体类型光谱，第 1、3 列分别展示了对应的天体类型的原始光谱和模拟了 CCD 坏像元后的光谱。

4.2.3 宇宙高能射线

在 CCD 相机接受光谱信息的同时，宇宙中的高能射线会有几率直接打在 CCD 相机的接

收端，使得某一随机波长处的流量强度极高。同样的，宇宙高能射线对光谱数据造成干扰的情况不多，但为了扩充数据和增加分类模型的泛化性和鲁棒性，我们通过将归一化后的光谱中某一随机波长处的流量置为 1 去模仿宇宙高能射线。增加模仿宇宙高能射线前后的光谱可视化如图 8。第 1、2、3 行展示了随机选取的恒星、星系、类星体天体类型光谱，第 1、4 列分别展示了对应的天体类型的原始光谱和模拟了宇宙高能射线后的光谱。

4.2.4 综合增强

在光谱数据的实际产生过程中，CCD 电子噪声、CCD 坏像元和宇宙高能射线的影响是同时存在的。为了使扩充的数据更加贴近真实情况，我们对光谱数据同时增加了这三种噪声的模拟。顺序依次是：原始光谱 -> 模拟 CCD 电子噪声 -> 模拟宇宙高能射线 -> 模拟 CCD 坏像元。综合考虑这三种噪声前后的光谱可视化如图 8。第 1、2、3 行展示了随机选取的恒星、星系、类星体天体类型光谱，第 1-5 列分别展示了对应的天体类型的原始光谱和依次模拟了 CCD 电子噪声、宇宙高能射线、CCD 坏像元后的光谱。

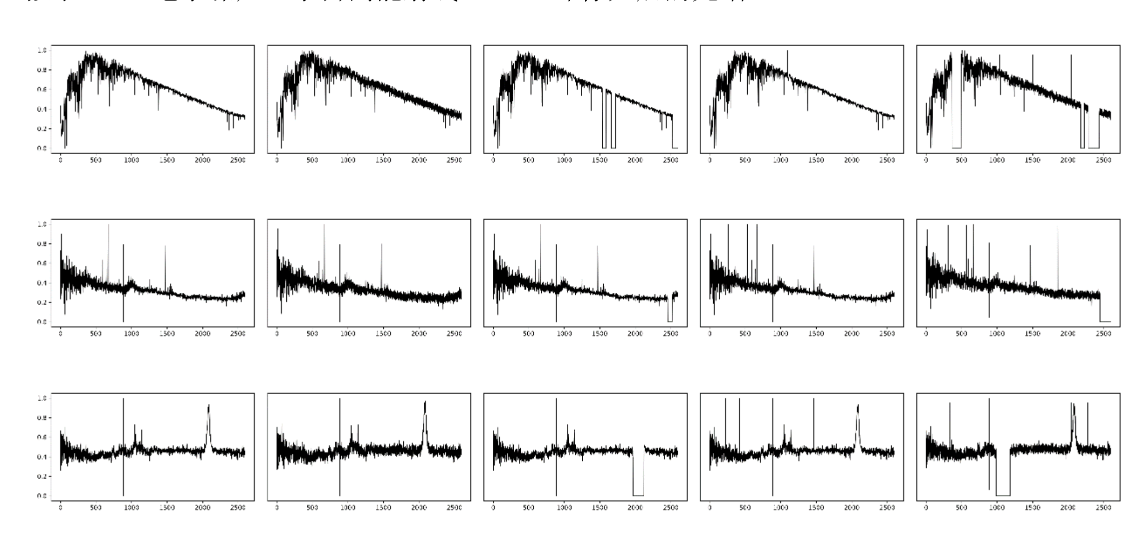


图 8 原始数据与增强后数据的可视化结果。第 1、2、3 行表示随机选取的恒星、星系、类星体天体类型光谱。第 1 列为原始光谱，第 2 列为模拟了 CCD 电子噪声后的光谱，第 3 列为模拟了 CCD 坏像元后的光谱，第 4 列为模拟了宇宙高能射线后的光谱，第 5 列为依次模拟 CCD 电子噪声、宇宙高能射线、CCD 坏像元后的光谱。

Fig.8 Visualization of raw data and augmented data. Rows 1, 2 and 3 represent the spectra of randomly selected stars, galaxies and quasars, respectively. Columns 1, 2, 3, 4 and 5 represent the original spectrum, the spectrum adding the electronic noise of CCD, the spectrum adding the bad pixels of CCD, the spectrum adding the cosmic high-energy ray, and the spectrum adding the electronic noise of CCD, the cosmic high-energy ray and the bad pixels of CCD, respectively.

4.3 模型训练

在训练卷积神经网络时使用 Adam 优化器更新网络权重[8]，使用交叉熵作为目标损失函数（loss）。卷积神经网络中所有卷积层权重采用 He 式均匀方差缩放初始化（He uniform

variance scaling initializer）[9]。神经网络共迭代训练 60 次，初始学习率设置为 0.01，分别在第 10、30、50 次时减小 10 倍。训练过程如图 x，训练的前 10 次，由于学习率很大，验证集上的 loss 很不稳定。随着学习率的减小和训练的增加，训练集和测试集上的 loss 都趋于稳定，并且慢慢变小。30 次训练后验证集 loss 几乎不再下降。为了防止过拟合，我们最终保留训练了 35 次的权重作为最终的模型。

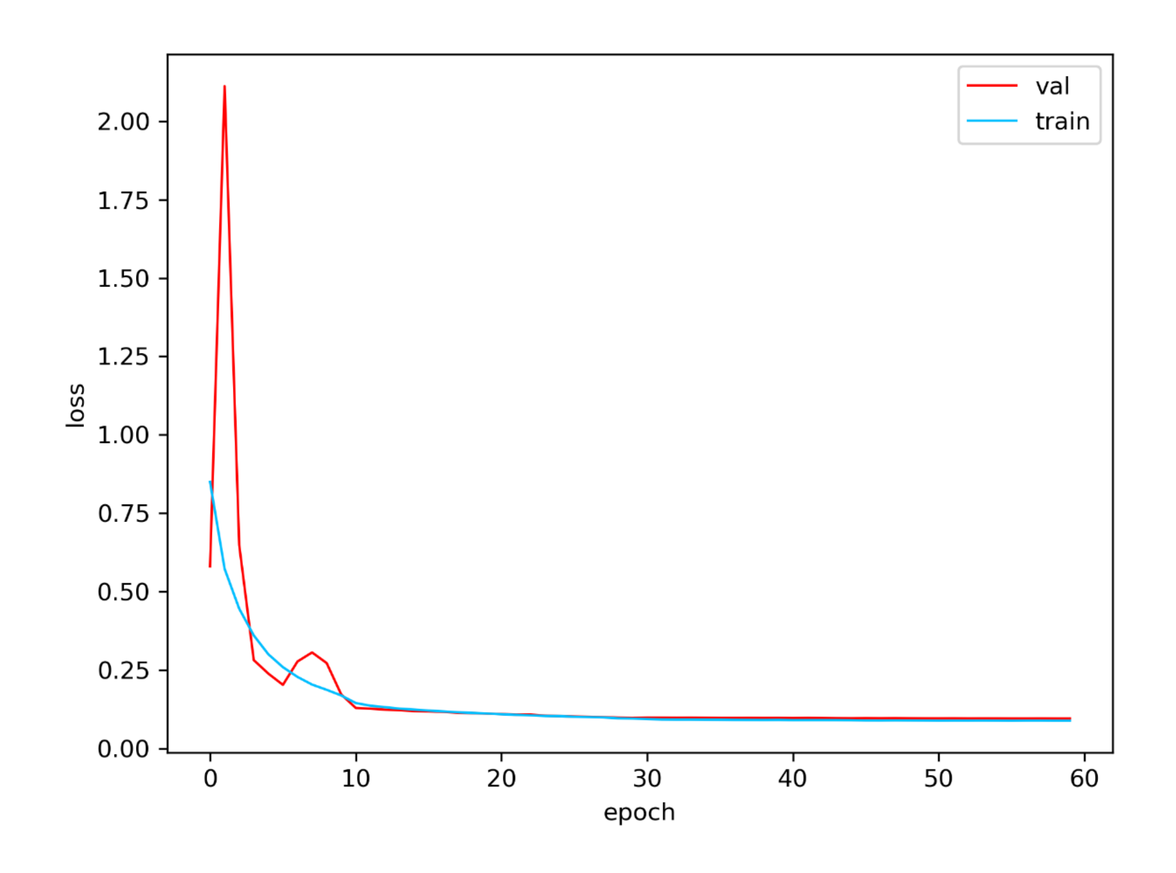


图 9 模型训练过程中的 loss 变化

Fig.9 Loss change progress during model training

5. 分类网络模型结果与关注区域可视化

我们将验证集数据在训练好的卷积神经网络上进行结果测试。验证集数据共 89877 条，恒星、星系和类星体数量占比与原始数据一致。

5.1 各类别 F1 分数及混淆矩阵

我们构建的卷积神经网络在验证数据进行分类，分类结果与验证数据的真实标签对比得到每一类的精确率、召回率和 F1 分数，如表 1，混淆矩阵如表 2。恒星类别天体光谱具有更强的特征，得到的恒星类别的精确率、召回率和 F1 分数均最高，F1 分数达到了 0.9987。星系与类星体的 F1 分数相近，但星系类型的召回率更高，类星体类型的精确率更高。

表 1 验证数据集上的精确率、召回率和 F1 分数

Tab.1 Precision, recall and F1-Score on validation data set

Categories	Precision	Recall	F1-Score
Star	0.9995	0.9980	0.9987
Galaxy	0.8593	0.9731	0.9127
Quasar	0.9231	0.9065	0.9147
Average	0.9258	0.9592	0.9420

混淆矩阵中可以看出，恒星类型预测错误的光谱中，160 条预测成了星系、18 条预测成了类星体，星系占比 89.89%，类星体占比 10.11%；星系类型预测错误的光谱中，25 条预测成了恒星、3 条预测成了类星体，恒星占比 89.29%，类星体占比 10.71%。类星体类型预测错误的光谱中，20 条预测成恒星，6 条预测成星系，恒星占比 76.92%，星系占比 23.08%。对于本文的卷积神经网络，同时考虑恒星、星系和类星体三种类型，恒星和星系具有更高的相似性；在类星体与恒星、星系相似性的比较中，类星体更加相似于恒星。

表 2 验证数据集上的混淆矩阵

Tab.2 Confusion matrix on validation data set

		Prediction		
		Star	Galaxy	Quasar
Reference	Star	88379	160	18
	Galaxy	25	1014	3
	Quasar	20	6	252

我们将本文模型的分类效果与传统机器神经网络方法、深度信念网络[3]和未改进版 DenseNet[4]的分类效果进行了对比，恒星、星系、类星体的平均指标如表 3。传统神经网络在出现了欠拟合问题，精确率、召回率和 F1 分数都显著低于其它方法，正是由于深层网络的梯度弥散或梯度爆炸所致。深度信念网络和 DenseNet 不同程度的解决了梯度弥散或梯度爆炸问题，精确率、召回率和 F1 分数有了明显提升。本文方法在 DenseNet 基础上进一步提升了各方面指标。

表 3 验证数据集上各方法平均分类指标

Tab.1 Average precision, recall and F1-Score of various methods on validation data set

methods	网络层数	Precision	Recall	F1-Score
传统神经网络	4	0.4260	0.2693	0.3299

深度信念网络	4	0.8166	0.8437	0.8299
DenseNet	36	0.9013	0.9294	0.9151
本文方法	36	0.9258	0.9592	0.9420

5.2 关注区域可视化

卷积神经网络拟合的光谱数据与天体类型间的关系是自主学习的，而天体类别是根据其物理意义进行划分的。我们希望神经网络自主学习到的是天体反映在光谱上具有物理意义的谱线特征，这样的神经网络具有更高的可解释性。为了检验本文训练得到的模型是否真正学到了可以反映天体类型的特征谱线，我们使用 Class Activation Mapping[10]（CAM）方法对神经网络模型在预测分类时关注的重点区域进行了可视化。CAM 是一种神经网络关注区域可视化方法，如图 10，具体流程为：1）获取神经网络预测类别 C；2）获取全局池化（GAP）后全连接层到类别 C 的权重 w；3）使用权重 w 加权求和全局池化前的所有特征图，得到求和后的特征图 M；4）归一化 M 并插值到原图大小。

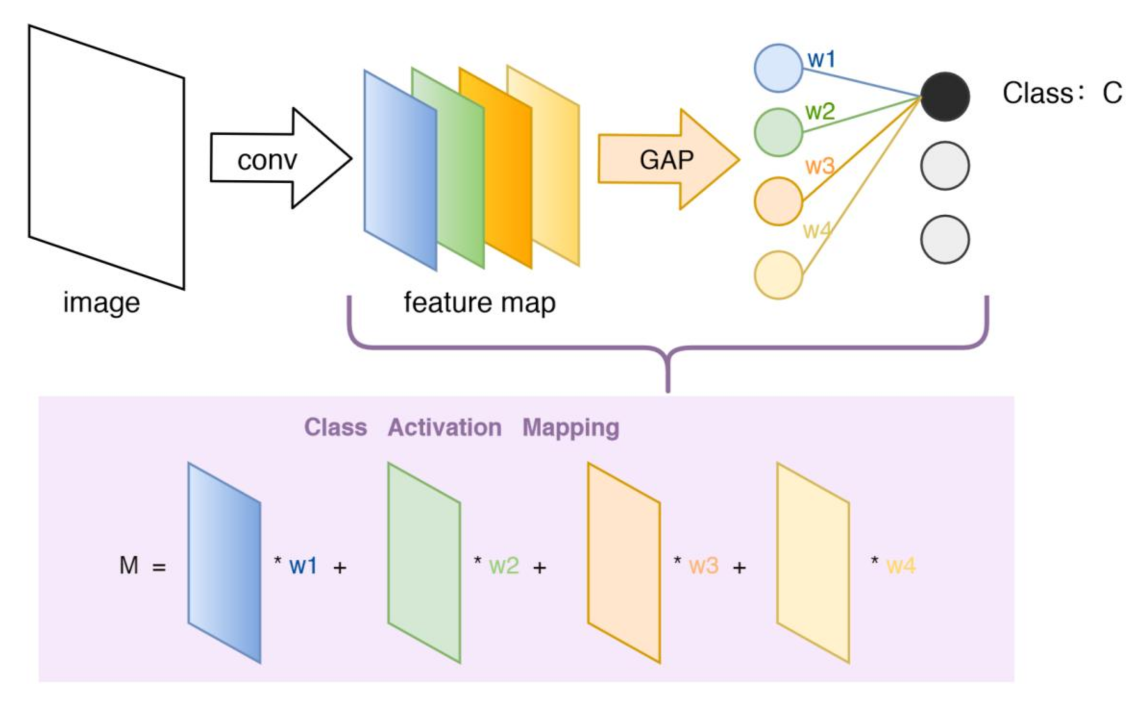
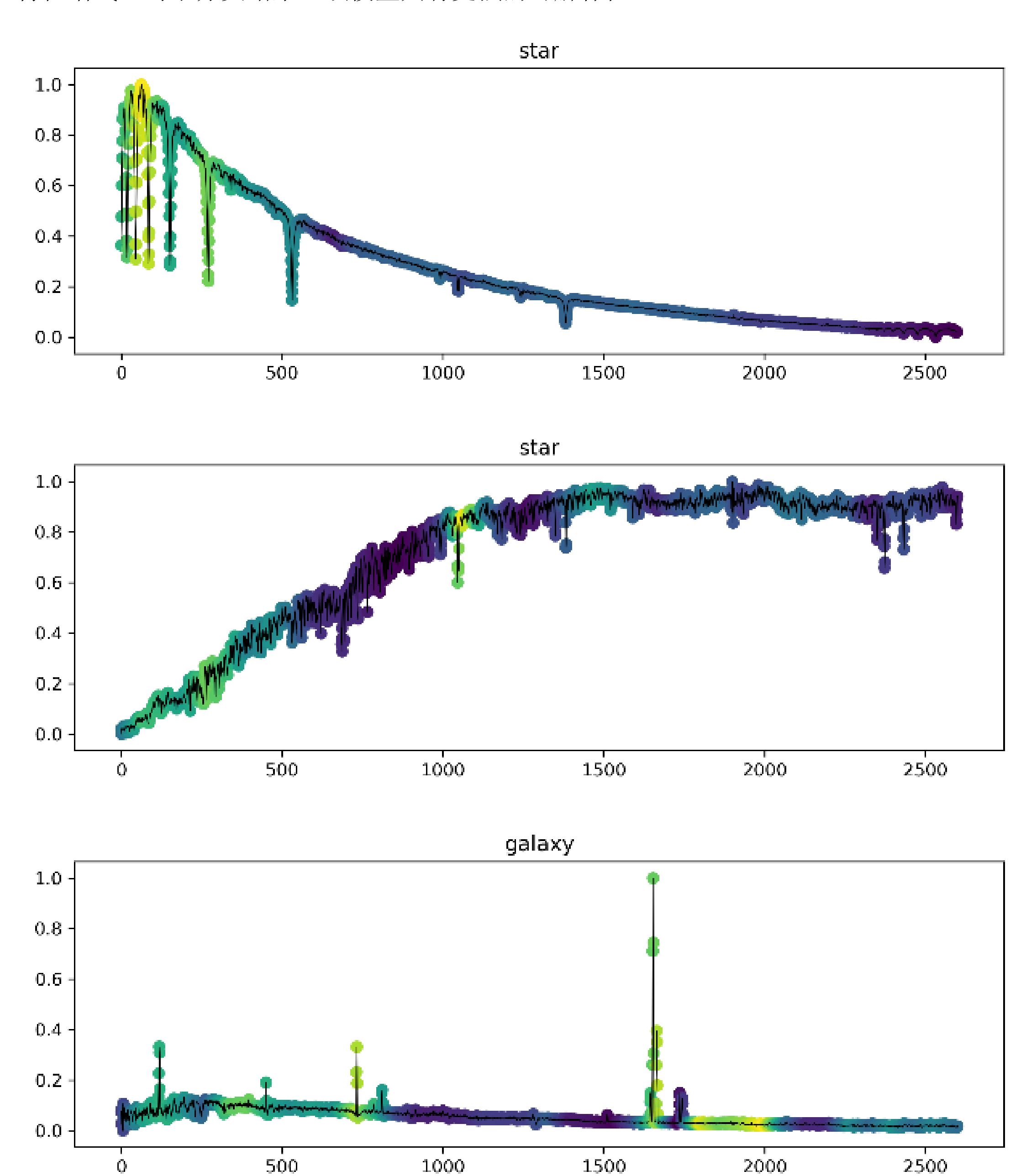


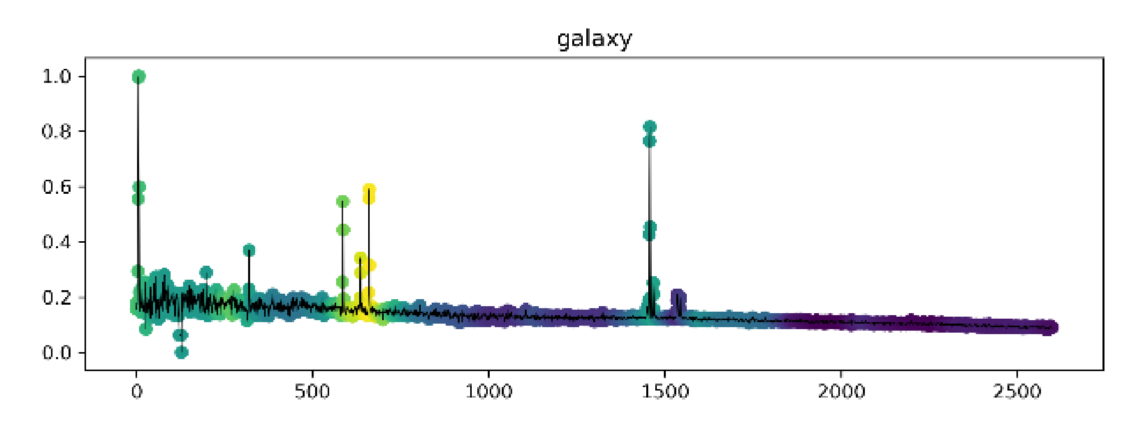
图 10 Class Activation Mapping

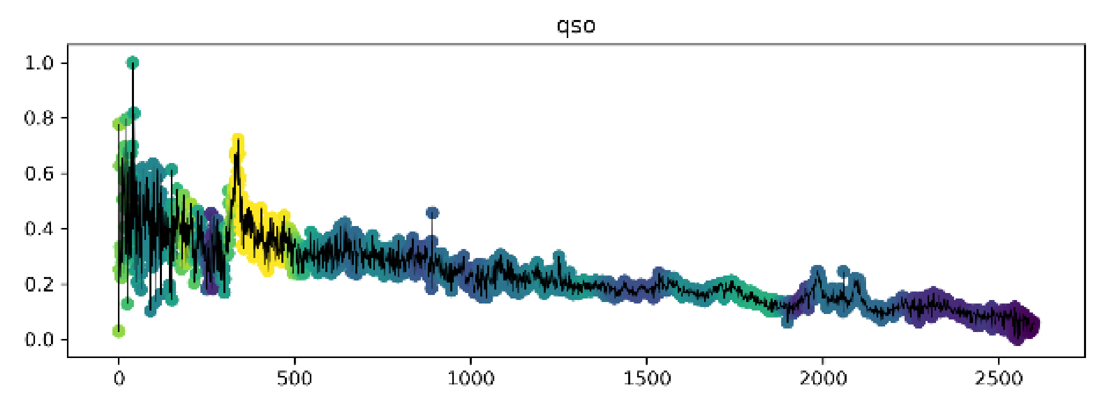
Fig.10 Class Activation Mapping

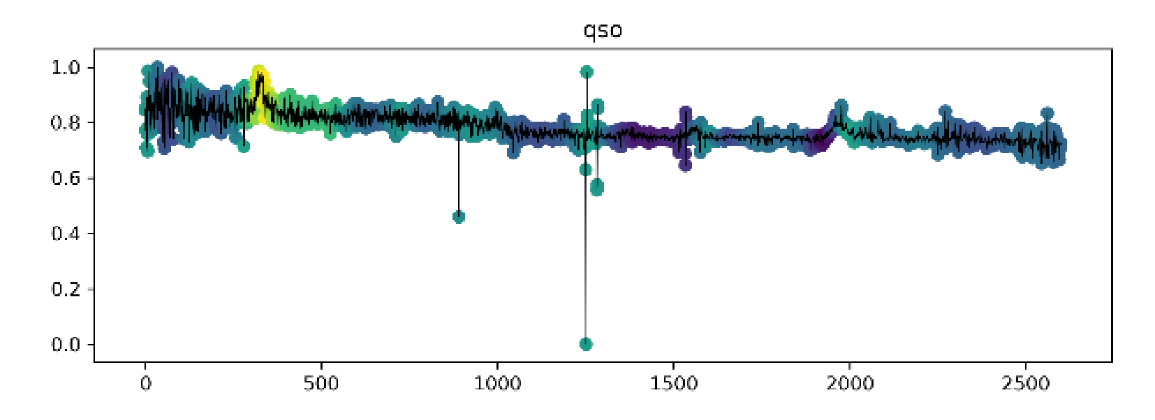
我们随机选择分类正确的恒星、星系、类星体和分类错误的光谱，将神经网络在不同波段处关注的强弱展示在原始光谱上，如图 11。第 1-6 行分别展示了 2 条随机选择的分类正确的恒星、星系和类星体光谱及神经网络关注的区域，第 7-10 行展示了分类错误的光谱及神经网络关注的区域。光谱中越黄、越亮的波段，代表神经网络更加关注这些区域在分类时的贡献；越蓝、越暗的波段，代表神经网络更加忽视这些区域在分类时的贡献。恒星与星系、

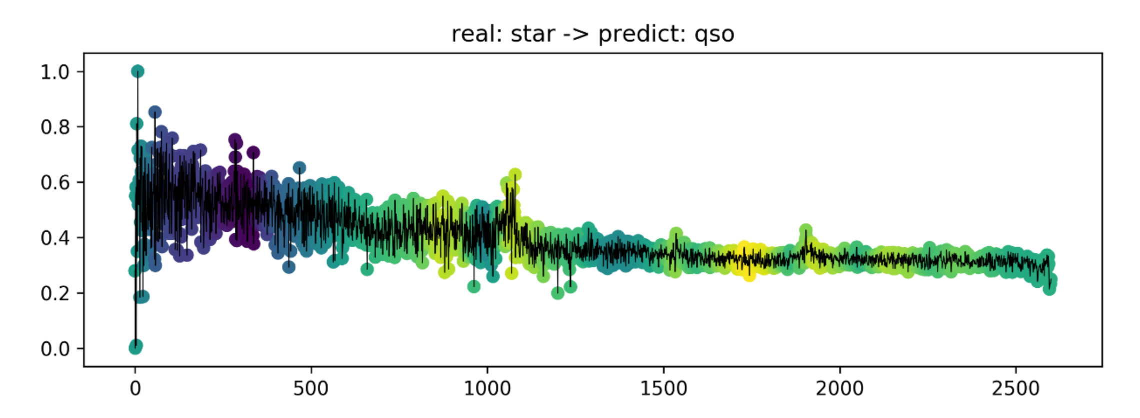
类星体在光谱上的显著差异，是其具有较强的氢元素的吸收线，这也是恒星大气相对于星系、类星体独有的特点。神经网络在对第 1、2 行的两条恒星光谱分类时，重点关注了氢元素的吸收线，和物理规律是一致的。神经网络在对第 3-6 行的星系、类星体分类时，即使存在一定红移，模型也重点关注了特定元素的发射线、发射带。神经网络在第 7-10 行错误分类时关注区域相对弥散，关键的特征谱线因为信噪比低等原因在光谱中不显著，神经网络找不到关键的特征谱线。本文训练的神经网络模型在进行分类时，可以定位到具有具体物理意义的特征谱线。对于分类结果，该模型具有更强的可解释性。




galaxy
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500


qso
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500


qso
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500


real: star -> predict: qso
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500

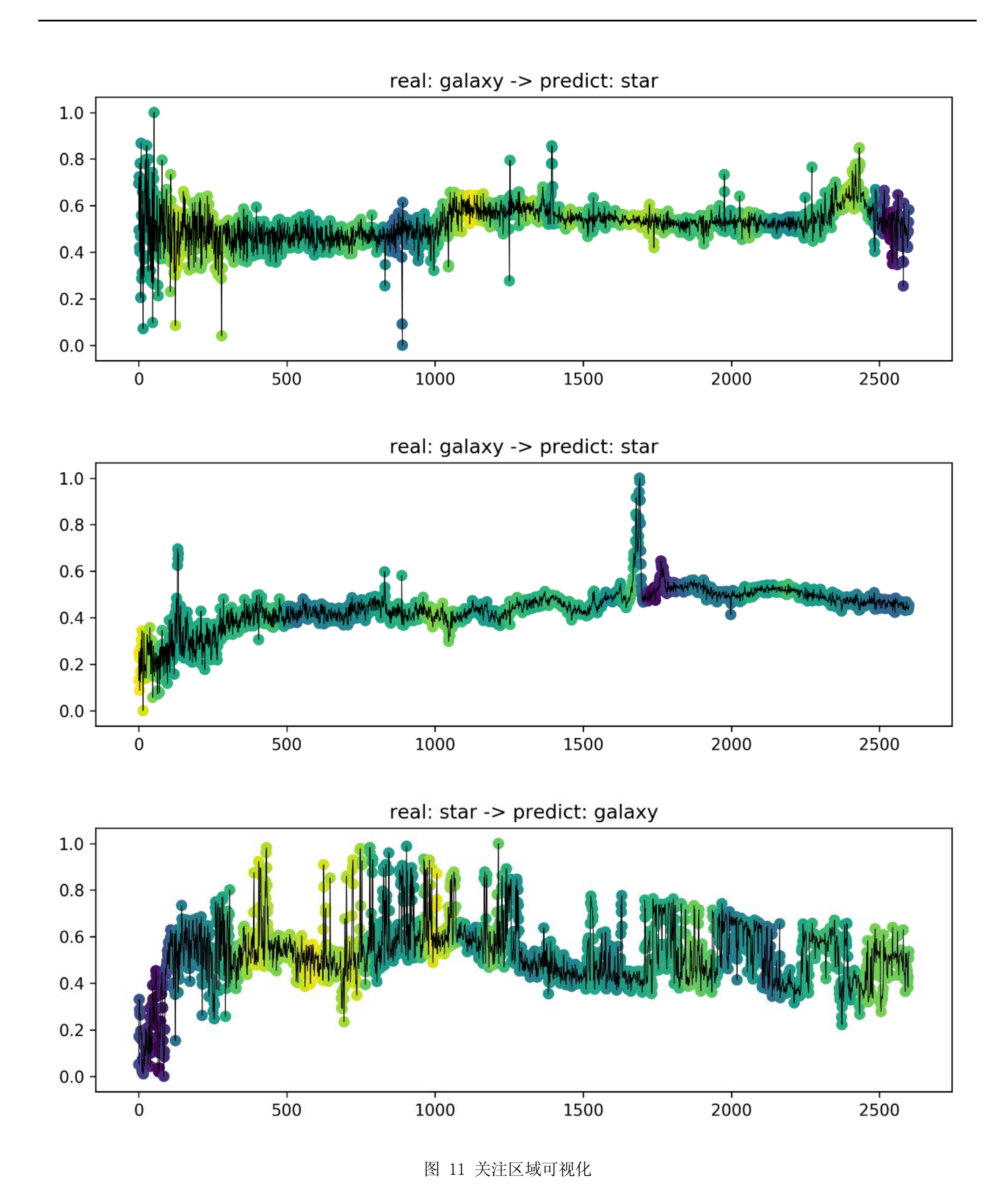


图 11 关注区域可视化

Fig.11 The visualization results of the regions of interest

6. 总结

本文基于在计算机视觉领域广泛应用的DenseNet网络结构，针对光谱数据进行修改，建立了适用于光谱数据的一维卷积神经网络模型解决天体光谱数据分类任务。通过模拟CCD电子噪声、CCD坏像元和宇宙高能射线平衡数据中不同天体类型的数量。本文的卷积神经网络模型在随机选取的89877条光谱测试中，恒星、星系、类星体的F1分数分别为0.9987、0.9127、0.9147。为了检验该卷积神经网络模型的可解释性，我们对神经网络模型在分类时

关注的区域进行了可视化。结果表明本文的卷积神经网络可以自主学习到各类天体对应的特征谱线，具有较强的可解释性。本方法被用于阿里云天池天文数据挖掘大赛——天体光谱智能分类，并在843支参赛队伍的3次数据评比中获得了2次第一、1次第三的成绩，证明了该模型在保证分类精度的同时具有极强的鲁棒性、泛化性，适用于光谱的自动分类。

致谢： 本论文得到中国虚拟天文台提供的数据资源和技术支持。感谢国家天文台-阿里云天文大数据联合研究中心、阿里云天池大数据众智平台对本项工作的支持。

参考文献

[1]潘儒扬，李乡儒. 基于深度学习技术的恒星大气物理参数自动估计[J]. 天文学报, 2016, 57(4).
Pan R, Li X. Stellar Atmospheric Parameterization Based on Deep Learning[J]. Acta Astronomica Sinica, 2016
[2]韩帅，李悦. 基于 BP 神经网络（自编码）的恒星大气物理参数估计[J]. 自动化与仪器仪表, 2016(9):230-231.
Han S, Li R. Stellar Atmospheric Parameterization Based on BP Neural Network (autoencode)[J]. Automation & Instrumentation, 2016
[3] 刘真祥，荣容，许婷婷，等. 基于深度信念网络的天体光谱自动分类研究[J]. 云南民族大学学报（自然科学版）, 2017(2).
Liu Z, Rong R, Xu T, et al. Automatic Classification of Star Spectra Based on the Deep Belief Network [J]. Journal of Yunnan Minzu University(Natural Sciences Edition), 2017.
[4] Huang G, Liu Z, Laurens V D M, et al. Densely Connected Convolutional Networks[J]. 2016.
[5] Ioffe S , Szegedy C . Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. 2015.
[6] Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks[J]. 2017.
[7] He K, Zhang X, Ren S, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification[J]. 2015.
[8] Kingma D , Ba J . Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.
[9] He K , Zhang X , Ren S , et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification[J]. 2015.
[10] Zhou B , Khosla A , Lapedriza A , et al. Learning Deep Features for Discriminative Localization[J]. 2015.

Classification of astronomical spectra based on DenseNet

Wang Qixun[12], Zhao Gang[12], Fan Zhou[1]

(1, Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101)

(2, School of Astronomy and Space Science, University of Chinese Academy of Sciences, Beijing 100049)

Abstract: Intelligent processing of astronomical spectra data is gradually shifting from traditional machine learning to deep learning, which mainly used the technology of computer vision. Based on DenseNet network structure, which is widely used in the field of computer vision, a one-dimensional convolution neural network model for spectral data is established to solve the classification task of celestial spectral data. The F1 scores of stars, galaxies and quasars are 0.9987, 0.9127 and 0.9147 respectively in the validated data set. The visualization results of the regions of interest in spectral classification show that the proposed model can learn the characteristic spectral lines of celestial bodies which has interpretability. This method is applied to the intelligent classification of celestial spectrum in Alibaba Cloud Tianchi Astronomical Data Mining Competition. We won the first prize two times and the third prize one time in 843 teams in three data evaluations, and the result proves that the model has robustness and generalization, and is suitable for automatic classification of spectra.

Key words: convolutional neural network; classification of spectra; data mining